W98–W103 *Nucleic Acids Research, 2015, Vol. 43, Web Server issue* *Published online 18 May 2015*
doi: 10.1093/nar/gkv511

Multidimensional gene search with Genehopper

Matthias Munz[1], Sascha Tönnies[1], Wolf-Tilo Balke[1] and Eric Simon[2,*]

[1]Institute for Information Systems, Technische Universität Braunschweig, Braunschweig 38106, Germany and [2]Target Discovery Research, Boehringer Ingelheim Pharma GmbH & Co. KG, Birkendorfer Str. 65, Biberach (Riss) 88397, Germany

Received January 31, 2015; Revised May 05, 2015; Accepted May 05, 2015

ABSTRACT

The high abundance of genetic information enables researchers to gain new insights from the comparison of human genes according to their similarities. However, existing tools that allow the exploration of such gene-to-gene relationships, apply each similarity independently. To make use of multidimensional scoring, we developed a new search engine named Genehopper. It can handle two query types: (i) the typical use case starts with a term-to-gene search, i.e. an optimized full-text search for an anchor gene of interest. The web-interface can handle one or more terms including gene symbols and identifiers of Ensembl, UniProt, EntrezGene and RefSeq. (ii) When the anchor gene is defined, the user can explore its neighborhood by a gene-to-gene search as the weighted sum of nine normalized gene similarities based on sequence homology, protein domains, mRNA expression profiles, Gene Ontology Annotation, gene symbols and other features. Each weight can be adjusted by the user, allowing flexible customization of the gene search. All implemented similarities have a low pairwise correlation (max r^2 = 0.4) implying a low linear dependency, i.e. any change in a single weight has an effect on the ranking. Thus, we treated them as separate dimensions in the search space. Genehopper is freely available at http://genehopper.ifis.cs.tu-bs.de.

INTRODUCTION

Emerging high-throughput technologies like next-generation sequencing (NGS) have led to a dramatic increase of descriptive and functional genetic information over the past decade, revealing gene properties such as gene and protein family, mRNA tissue distribution, gene functional or pathway membership. Many very powerful tools have evolved to explore this information, e.g. Ensembl, RefSeq, Homologene, UniProt, InterPro, Gene Ontology (GO), KEGG, REACTOME, GTExPortal, THE HUMAN PROTEIN ATLAS, or even meta-tools like STRING or MSigDB. Further processing of these properties into gene similarities enables the unbiased exploration of inter-gene relationships, representing a useful application in target discovery research.

Since in the present study we focus on human coding genes we use the term ‘gene’ as a generic term for a gene locus and the corresponding gene products, i.e. transcripts and proteins.

Several computational tools exist for finding gene relationships. The UCSC GeneSorter (1) (available at: https://genome.ucsc.edu/cgi-bin/hgNear) ranks genes from human, rat and mouse based on the similarity of expression profiles of the GNF Gene Expression Atlas, GO terms (2), the proximity in genome, the gene name and both structural and functional protein similarities. It provides additional filtering options for restricting the gene corpus, e.g. to a certain chromosomal position. EvoCor (3) (available at: http://pilot-hmm.vbi.vt.edu) ranks genes according to putative functional linkages with a gene of interest. This functional similarity is predicted from evolutionary information, i.e. sequence homology and gene expression data. It incorporates genes from all eukaryotic species in the NCBI database. LineUp (4) (available at: http://caleydo.org/projects/lineup) is another, generic approach which uses bar charts to rank items in an interactive way. It allows multiple heterogenous attributes, i.e. has no restrictions on certain scales or semantics. However, compared to the UCSC GeneSorter and EvoCor it is not query based, i.e. does not rank objects according to the similarity to a query (e.g. a gene) and therefore only provides limited possibilities for finding gene-to-gene relationships. Instead, it can rather be used as a tool in the field of gene prioritization.

In the tools described above the number of genetic similarities is small and/or each similarity is used independently without making use of multidimensional scoring. In this paper, we present a new search engine with a focus on human genes that incorporates multiple gene similarities into a weighted linear model. The weighted model allows the user to combine these similarities into a single ranking score ac-

Downloaded from http://nar.oxfordjournals.org/ at Michigan State University on November 20, 2015

[*]To whom correspondence should be addressed. Tel: +49 7351 54 94883; Fax: +49 7351 83 94883; Email: eric.simon@boehringer-ingelheim.com
Present address: Matthias Munz, Molecular Tumor Pathology, Institute of Pathology, Charité, Berlin 10117, Germany and Computational Modelling in Medicine, Institute of Theoretical Biology, Humboldt-Universität, Berlin 10099, Germany.

© The Author(s) 2015. Published by Oxford University Press on behalf of Nucleic Acids Research.
This is an Open Access article distributed under the terms of the Creative Commons Attribution License (http://creativecommons.org/licenses/by/4.0/), which permits unrestricted reuse, distribution, and reproduction in any medium, provided the original work is properly cited.

cording to the user's needs. Additionally, we put a strong emphasis on the speed and accuracy of the web based search and retrieval front-end. To integrate our tool into the landscape of public available resources, the search results links to external web sites such as Ensembl or PubMed. We chose the name Genehopper to clarify the ability of easy and fast navigating through the corpus of human genes.

MATERIALS AND METHODS

The architecture of Genehopper consists of a data integration and a data application component (as seen on the left and right of Figure 1, respectively). Data from various public databases is extracted, transformed into the internal data model and finally loaded into a local MySQL database (ETL process). The ETL procedure is implemented in Perl and fully automated to allow regular updates.

Data sources

Most of the data has been extracted from Ensembl (5) version 79 via Ensembl Perl API or SQL. From the Ensembl Core database we got information on positional data of genes, exons and transcripts, Ensembl identifier mappings to other databases such as InterPro, UniProt, RefSeq, EntrezGene, Gene Ontology, InterPro as well as external reference data including gene names, descriptions and synonyms. From the Ensembl Compara database we got information about paralogous and orthologous genes and the corresponding amino acid sequence based homologies. From the Ensembl Variation database (6) we got information concerning non-structural genetic variants including dbSNP identifier, synonyms and variant related citations. Data extraction from Ensembl Core and Ensembl Compara was done by using the Ensembl Perl API. For data from the Ensembl Variation database we chose SQL due to performance reasons. Abstracts of the cited publications were fetched from PubMed with the tool E-Utilities (documentation available at: http://www.ncbi.nlm.nih.gov/books/NBK25497). Protein feature information was fetched from the manually curated UniProtKB/Swiss-Prot database.

All semantic gene information has been compiled into a single MySQL table. For this table a full-text index was created to associate user query terms to target genes. An exemplary database table is shown in Supplementary Figure S1.

Search engine - Concept and implementation

The web front-end application is implemented in Java and Javascript by utilizing the Play! Framework (available at: https://www.playframework.com), jQuery (available at: http://jquery.com) and Twitter Bootstrap (available at: http://getbootstrap.com). The term-to-gene search was realized using a two-step approach: (i) the user query terms are matched against the full-text index described above giving a set of hit genes that match the query. (ii) These hits are then ranked by the type of match e.g. a match to the HGNC symbol is ranked higher compared to a match to one of the synonyms or a substring match.

The gene-to-gene search is based on the ranked multidimensional gene similarities. In order to calculate an overall ranking score for each gene we use a linear weighed sum model according to:

$$S = f(\vec{x}) = w_1 x_1 + w_2 x_2 + \cdots + w_n x_n \qquad (1)$$

x denotes the similarity vector of size n for a pair of anchor gene and result gene. Each similarity x_i is multiplied with a weight w_i to adapt the ranking. Result genes are ranked in descending order according to their overall ranking score which corresponds to the weighted sum of the pairwise similarities between the anchor gene and all other genes. To facilitate interpretation of the similarity values and the overall ranking score we constrain the normalization of the weights and similarities according to:

$$\sum\nolimits_{i=1}^{n} w_i = 1 \qquad (2)$$

$$0 \le w_i \le 1$$

$$0 \le x_i \le 1$$

To improve the search performance we pre-calculated the similarities as part of the automated ETL procedure. Only the similarity values above zero are kept and stored to minimize the database size. Moreover for both the term-to-gene and the gene-to-gene search the result lists of the 100 most frequent queries are cached using the EHCache (available at: http://ehcache.org) which is part of the Play! framework.

Gene similarities

Altogether we selected nine different gene and protein similarities which have been scored according to five different similarity measures (see Table 1) to be most suitable: (1) Accordingly, S_{HOM} corresponds to the protein sequence identity according to the Ensembl Compara database. For all other genes the homology similarity value is set to zero.

(2) S_{IPD} represents the similarity of protein domains and is derived from the presence or absence of InterPro protein domains in the corresponding Ensembl gene product entry. We applied cosine over the Jaccard or Dice measure because it produces higher similarity values for sets with very different sizes and a relatively big intersection.

(3) The similarity of publications related to sequence variants S_{VP} is calculated from publications that are associated with two given genes describing sequence variants in the gene loci. In most cases, the publications correspond to genome wide associations (GWA) of single nucleotide polymorphisms with phenotypic traits (GWAS). Again we applied cosine for the same reasons as mentioned before. The publication-variant associations were extracted from Ensembl Variation database which Ensembl in turn collected from dbSNP (7), EPMC (available at: http://europepmc.org) and UCSC (8).

(4) The similarity of protein features S_{SPF} was calculated from a binary vector of 24 protein features using Cosine again. These features are derived from the UniProtKB/Swiss-Prot database and include binary flags for subcellular localization, protein family membership and role in human diseases according to OMIM (Online

Downloaded from http://nar.oxfordjournals.org/ at Michigan State University on November 20, 2015

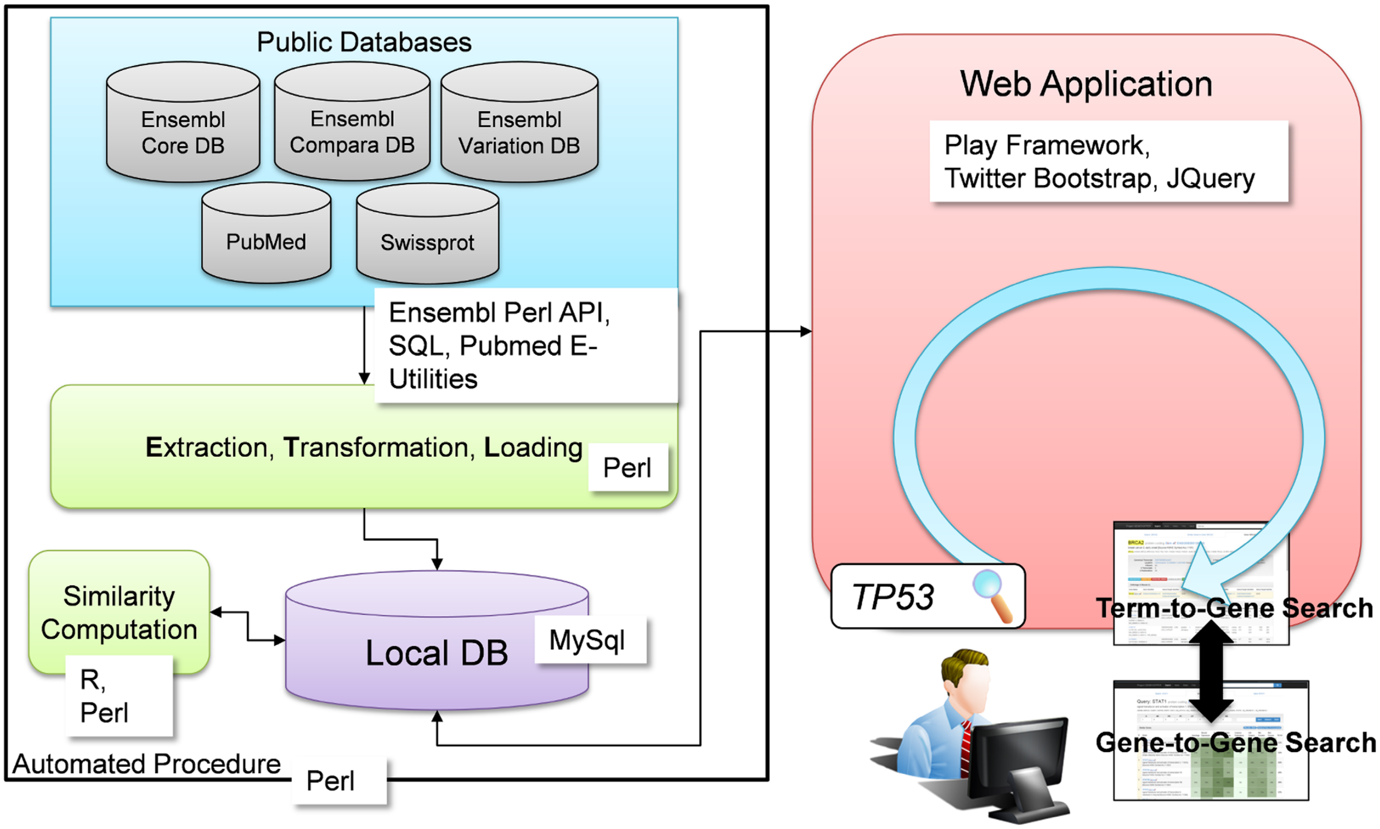


Figure 1. Data from various public databases is fetched to create a local and integrated database (left component). The web application of Genehopper is used to access the data by providing the search types term-to-gene search and gene-to-gene search (right component). The workflow on the web front-end is initiated by a text query, e.g. *TP53*.

Table 1. Gene similarities used in the gene-to-gene search and the respective similarity measure

	Similarity	Data source	Similarity measure
1.	Homology (S_{HOM})	Ensembl Compara	Sequence identity
2.	InterPro protein domain (S_{IPD})	Ensembl Core	Cosine
3.	Gene variant related publications (S_{VP})	Ensembl Variation	Cosine
4.	Swiss-Prot protein feature (S_{SPF})	UniProtKB/Swiss-Prot	Cosine
5.	GO cellular component (S_{CC})	Ensemble Core	Resnik-BMA
6.	GO molecular function (S_{MF})	Ensemble Core	Resnik-BMA
7.	GO biological process (S_{BP})	Ensemble Core	Resnik-BMA
8.	Normal tissue expression profile (S_{NEX})	Human Protein Atlas	Spearman
9.	HUGO gene symbol (S_{HGS})	HGNC	Prefix distance

Mendelian Inheritance in Man, http://omim.org/) (detailed list in Supplementary Table S1). Generally, we mapped each Ensembl gene to the corresponding UniProtKB/Swiss-Prot entry by cross-references of both databases to each other.

(5–7) Similarities of GO terms S_{CC}, S_{MF}, S_{BF} have been investigated extensively by others. Accordingly, we applied Resnik as single GO term similarity measure (9–12). As combination method to derive the overall similarity for two term sets we chose Best-Match Average. We used the R implementation in the GOSemSim package (13).

(8) For the similarity of gene expression in normal tissue S_{NEX} we used a recently published mRNA Seq panel of 27 normal tissues (14,15). As similarity measure we used Spearman rank correlation coefficient.

(9) The last similarity of gene symbol is computed from HUGO gene symbols from HGNC (16). As similarity measure we chose the prefix distance (Supplementary Figure S2).

RESULTS AND DISCUSSION

Our goal with Genehopper was to provide a tool which combines multidimensional similarities of genes to systematically explore the neighborhood of human genes. Therefore, we subdivide the workflow on the web-based user interface into two steps. First, an initial anchor gene search by one or more keywords is performed (term-to-gene search). Next, the gene selected from the result list serves as the anchor gene for the multidimensional gene-to-gene search. In both steps, every gene that is shown can be loaded in a gene details view.

Downloaded from http://nar.oxfordjournals.org/ at Michigan State University on November 20, 2015

Term-to-gene search

For selecting an anchor gene via the term-to-gene search, Genehopper is not restricted to certain input terms. Besides gene symbols and identifiers of Ensembl, UniProt, EntrezGene, RefSeq, Genehopper can also find genes by publication or variant identifiers, even unspecific vocabulary is handled. Typically, human users do not remember long and cryptical gene identifiers or database accession numbers from Ensembl, Refseq, UniProt & Co. In most cases, they use the gene symbol or a substring for searching. This could be demonstrated in a survey of user queries that have been sent to a gene search tool (17). Consequently, our term-to-gene ranking search is superior to many other search engines which require exact input queries in many cases (Supplementary Tables S2 and S3). Due to the optimized full-text index and caching mechanism the retrieval speed of the term-to-gene search is highly competitive with other search engines (Supplementary Figure S3).

The result page of term-to-gene search contains information about the result genes, including a gene symbol, hyperlinks to external references, a single-line description, synonyms and identifiers. To track how a result gene was derived from the query, information about the matched data model attributes is listed as well. Depending on the type of query match, a result gene is ranked higher, i.e. is more relevant to the user, or lower, i.e. is less relevant to the user. Strings or substrings in the result list that match the query terms are backed with yellow.

Gene-to-gene search

Figure 2 shows the results page of the gene-to-gene search for one example, i.e. *TP53*. Genes in the result list are sorted descending according to their weighted pairwise similarities, i.e. ranking scores with respect to the anchor gene *TP53*.

The tool allows flexible customization of the gene search according to specific use cases by setting a pre-specified weight for each single similarity in the upper row of the search result panel. To facilitate the handling of the weight profiles, which by virtue of the restriction that the sum of all weights is one would contain odd values in the default case for nine similarities, each weight is multiplied by the number of similarities and displayed in percentage, leading to default values of one and 100 percent respectively. Adjusted weight profiles can be used to re-rank the result set (button on the right of weight settings) or saved to use them for future searches by setting a HTTP Cookie (buttons on the left of weight settings).

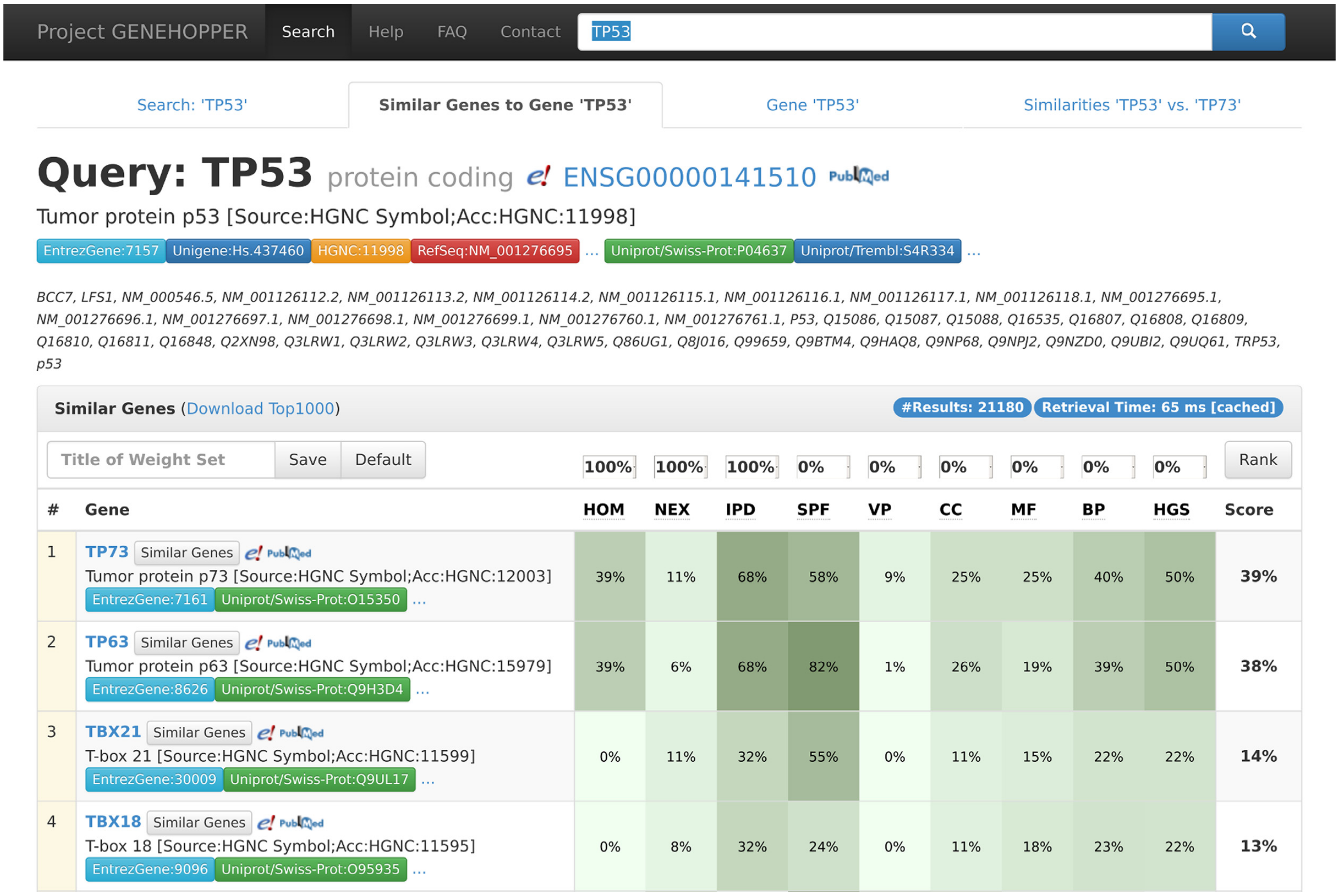
The similarity values are also shown in percentage and backed with a linear white-to-green color gradient. In contrast to other search engines, we employ a multidimensional similarity ranking with normalized ranking scores to illustrate the gene-to-gene relationships. The linear weighted similarity sum model allows a highly flexible ranking approach which could be useful for other bioinformatics applications, i.e. gene network construction, target pre-selection or candidate gene prioritization. EvoCor applies a graded sorting of the result genes, i.e. the result list is first sorted by sequence similarity and then by expression similarity. The UCSC GeneSorter only allows a sorting by a single similarity. When looking at individual gene similarities, UCSC GeneSorter handles all GO terms as a single set and measures the term overlap between genes. To expose a more finely granulated sight on the information in the Gene Ontology, we handle the three sub-ontologies molecular function, biological process and cellular component as separate similarities, not considering, e.g. process/function links between the sub-ontologies. LineUp allows the weighting of single attributes, but doesn't take advantage of similarity measures due to its non-query based architecture.

Moreover, we provide normalized (interval [0;1]) similarities and overall ranking scores in percentage quotation. This allows the user not only to infer similarity by assuming equal distance between every position, but to interpret each gene by a proper similarity value. We also evaluated the suitability of a machine learning algorithm for the personalization of the ranking (17). However, we did not follow this approach for the current study because different users prefer different rankings and even for a single user the preferred ranking can vary depending on the task.

To reduce the number of dimensions, we analysed and evaluated additional similarities and measures prior to the actual deployment of Genehopper. For this purpose we did a pairwise correlation of all similarities with all similarity values greater than zero (input data sizes in Supplementary Table S4). Table 2 shows the pairwise Pearson and Spearman correlation between the similarities (P-value $< 2.2\mathrm{e}{-16}$ under the null hypothesis that the correlation is 0). Generally, both correlation coefficients have a very good agreement. This indicates that linear and monotonic trends are equally distinct. Regarding the correlation of similarities, the poorest correlation is observed between expression S_{NEX} and all other similarities (Pearson: [−0.04;0.08], Spearman: [−0.07;0.08]) except with S_{HOM} (Pearson: 0.27, Spearman: 0.21) as well as gene variant publication S_{VP} (Pearson: [−0.04;0.07], Spearman: [−0.16;0.03]) except with S_{HOM}, S_{IPD}, S_{HGS} (Pearson: [0.26;0.53], Spearman: [0.03;0.48]). Homology S_{HOM} has a relatively high positive correlation with all other similarities (correlation coefficient ≥ 0.21) except with S_{HGS} (Pearson: −0.17, Spearman: −0.15) and a maximum of 0.53 with S_{VP}. For the other similarities, a somewhat higher correlation (Pearson: [0.16, 0.63], Spearman: [0.12, 0.63]) could be measured. Obviously, all similarities are not completely independent from each other. However, due to the relatively low overall correlations (max $r^2 = 0.4$), we decided not to merge any of the similarities and to treat all of them as separate dimensions in the search space. The low correlations also imply that a change in weights has an effect to the ranking.

In order to gather a deeper knowledge of gene-to-gene relationships, we provide raw data from which the similarities were calculated by linking to a similarities page for each pair of anchor gene and result gene (link on similarity values). For further analysis with other tools the raw data for a pair of genes as well as the similarity values of the top 1000 result genes can be downloaded as text files.

Downloaded from http://nar.oxfordjournals.org/ at Michigan State University on November 20, 2015



Query: TP53 protein coding ENSG00000141510

Tumor protein p53 [Source:HGNC Symbol;Acc:HGNC:11998]

EntrezGene:7157 Unigene:Hs.437460 HGNC:11998 RefSeq:NM_001276695 … Uniprot/Swiss-Prot:P04637 Uniprot/Trembl:S4R334 …

BCC7, LFS1, NM_000546.5, NM_001126112.2, NM_001126113.2, NM_001126114.2, NM_001126115.1, NM_001126116.1, NM_001126117.1, NM_001126118.1, NM_001276695.1, NM_001276696.1, NM_001276697.1, NM_001276698.1, NM_001276699.1, NM_001276760.1, NM_001276761.1, P53, Q15086, Q15087, Q15088, Q16535, Q16807, Q16808, Q16809, Q16810, Q16811, Q16848, Q2XN98, Q3LRW1, Q3LRW2, Q3LRW3, Q3LRW4, Q3LRW5, Q86UG1, Q8J016, Q99659, Q9BTM4, Q9HAQ8, Q9NP68, Q9NPJ2, Q9NZD0, Q9UBI2, Q9UQ61, TRP53, p53

Similar Genes (Download Top1000) #Results: 21180 Retrieval Time: 65 ms [cached]

#	Gene	HOM	NEX	IPD	SPF	VP	CC	MF	BP	HGS	Score
		100%	100%	100%	0%	0%	0%	0%	0%	0%	
1	**TP73** Similar Genes Tumor protein p73 [Source:HGNC Symbol;Acc:HGNC:12003] EntrezGene:7161 Uniprot/Swiss-Prot:O15350 …	39%	11%	68%	58%	9%	25%	25%	40%	50%	**39%**
2	**TP63** Similar Genes Tumor protein p63 [Source:HGNC Symbol;Acc:HGNC:15979] EntrezGene:8626 Uniprot/Swiss-Prot:Q9H3D4 …	39%	6%	68%	82%	1%	26%	19%	39%	50%	**38%**
3	**TBX21** Similar Genes T-box 21 [Source:HGNC Symbol;Acc:HGNC:11599] EntrezGene:30009 Uniprot/Swiss-Prot:Q9UL17 …	0%	11%	32%	55%	0%	11%	15%	22%	22%	**14%**
4	**TBX18** Similar Genes T-box 18 [Source:HGNC Symbol;Acc:HGNC:11595] EntrezGene:9096 Uniprot/Swiss-Prot:O95935 …	0%	8%	32%	24%	0%	11%	18%	23%	22%	**13%**

Figure 2. Result page of the gene-to-gene search with the exemplary query gene TP53. Each line in the result panel corresponds to a gene for which eight similarity scores and a ranking score is displayed. The ranking can be adjusted by reconfiguring the weights in the panel above the result list. Additional information about each similarity can be displayed by clicking on the similarity title.

Table 2. (Grey) Pearson and (white) Spearman correlations for each pair of the nine gene similarities

	S_{HOM}	S_{IPD}	S_{VP}	S_{SPF}	S_{CC}	S_{MF}	S_{BP}	S_{NEX}	S_{HGS}
S_{HOM}		0.43	0.53	0.28	0.38	0.39	0.42	0.27	-0.17
S_{IPD}	0.47		0.43	0.5	0.55	0.63	0.6	0.08	0.43
S_{VP}	0.48	0.37		0.05	0.04	0.07	-0.01	-0.04	0.26
S_{SPF}	0.32	0.51	0.02		0.48	0.19	0.25	0.08	0.31
S_{CC}	0.40	0.52	-0.04	0.49		0.16	0.3	0.07	0.33
S_{MF}	0.42	0.63	-0.01	0.16	0.15		0.26	0.03	0.33
S_{BP}	0.44	0.55	-0.16	0.18	0.22	0.2		0.06	0.48
S_{NEX}	0.21	0.08	-0.07	0.08	0.06	0.02	0.05		0.06
S_{HGS}	-0.15	0.43	0.03	0.13	0.14	0.12	0.19	0.01	

CONCLUSION AND OUTLOOK

Genehopper is a high performant search engine that extends the set of tools to explore human gene-to-gene relationships by a large number of gene similarities and the possibility to combine them into a multidimensional scoring model which facilitates the interpretation and increases the information content of rankings. Practically, the tool represents a powerful entry point to explore the molecular data and neighborhood of a given human gene or protein within the tool and by using the linked out resources. This will be useful for example to quickly evaluate functional similarities within larger gene families or to get a first insight into the function of genes where almost nothing is known from the literature.

To ensure relevance and reliability, the automated ETL procedure will be used to update the database regularly according to the update frequency of Ensembl (approximately twice per year) and also to include additional similarities which have not yet been included in the weighted ranking approach, i.e. curated pathway information, protein-protein interactions and genomic coordinates. Additionally, it would be very useful to provide a human readable summary of the similarity data, i.e. GO and InterPro names in addition to IDs, a visualization of the expression data ('Electronic Northern Blot') and the subcellular localization as well as the functional gene and protein information. Another potential direction for the future development of Genehopper is represented by personalized ranking. This could allow, i.e. optimized adjustment of the similarity weights for individual users. Furthermore, this would open the possibility to build up social network platforms for scientific groups which work on similar molecular targets, pathways or diseases. We tested already machine learning tools to predict the weight preferences of individual users

Downloaded from http://nar.oxfordjournals.org/ at Michigan State University on November 20, 2015

(17). However, such an approach will heavily rely on a large user community.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

ACKNOWLEDGEMENTS

Special thanks to Ramona Schmid and Ogsen Garbrielyan who supported this work with very useful advice and discussions and Holger Klein who was the driving force behind the initiation of the publication process.

FUNDING

Boehringer Ingelheim Pharma GmbH & Co. KG. Funding for open access charge: #43050340 Cooperation Agreement between Boehringer Ingelheim Pharma GmbH & Co. KG & Institute for Information Systems, Technische Universität Carolina Wilhelmina zu Braunschweig.

Conflict of interest statement. None declared.

REFERENCES

1. Kent,W.J., Hsu,F., Karolchik,D., Kuhn,R.M., Clawson,H., Trumbower,H. and Haussler,D. (2005) Exploring relationships and mining data with the UCSC Gene Sorter. *Genome Res.*, **15**, 737–741.
2. Ashburner,M., Ball,C.A., Blake,J.A., Botstein,D., Butler,H., Cherry,J.M., Davis,A.P., Dolinski,K., Dwight,S.S., Eppig,J.T. *et al.* (2000) Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nat. Genet.*, **25**, 25–29.
3. Dittmar,W.J., McIver,L., Michalak,P., Garner,H.R. and Valdez,G. (2014) EvoCor: a platform for predicting functionally related genes using phylogenetic and expression profiles. *Nucleic Acids Res.*, **42**, W72–W75.
4. Gratzl,S., Lex,A., Gehlenborg,N., Pfister,H. and Streit,M. (2013) LineUp: visual analysis of multi-attribute rankings. *IEEE Trans. Visual. Comput. Graph.*, **19**, 2277–2286.
5. Flicek,P., Amode,M.R., Barrell,D., Beal,K., Billis,K., Brent,S., Carvalho-Silva,D., Clapham,P., Coates,G., Fitzgerald,S. *et al.* (2014) Ensembl 2014. *Nucleic Acids Res.*, **42**, D749–D755.
6. Chen,Y., Cunningham,F., Rios,D., McLaren,W.M., Smith,J., Pritchard,B., Spudich,G.M., Brent,S., Kulesha,E., Marin-Garcia,P. *et al.* (2010) Ensembl variation resources. *BMC Genomics*, **11**, 293.
7. Sherry,S.T., Ward,M.H., Kholodov,M., Baker,J., Phan,L., Smigielski,E.M. and Sirotkin,K. (2001) dbSNP: the NCBI database of genetic variation. *Nucleic Acids Res.*, **29**, 308–311.
8. Meyer,L.R., Zweig,A.S., Hinrichs,A.S., Karolchik,D., Kuhn,R.M., Wong,M. *et al.* (2012) The UCSC Genome Browser database: extensions and updates 2013. *Nucleic Acids Res.*, **41**, D64–D69.
9. Resnik,P. (1999) Semantic similarity in a taxonomy: an information-based measure and its application to problems of ambiguity in natural language. *J. Artif. Intell. Res.*, **11**, 95–130.
10. Cho,Y.-R. (2013) Filtering protein-protein interactions by integration of ontology data. In: Elloumi,M and Zomaya,AY (eds). *Biological Knowledge Discovery Handbook*. John Wiley & Sons, Inc., Hoboken, NJ, doi:10.1002/9781118617151.ch04.
11. Pesquita,C., Faria,D., Falcão,A.O., Lord,P., Couto,F.M. and Bourne,P.E. (2009) Semantic similarity in biomedical ontologies. *PLoS Comput. Biol.*, **5**, doi:10.1371/journal.pcbi.1000443.
12. Alvarez,M.A., Qi,X. and Yan,C. (2011) GO-based term semantic similarity. In: Wong,W, Liu,W and Bennamoun,M (eds). *Ontology Learning and Knowledge Discovery Using the Web*. IGI Global, doi:10.4018/978-1-60960-625-1.ch009.
13. Yu,G., Li,F., Qin,Y., Bo,X., Wu,Y. and Wang,S. (2010) GOSemSim: an R package for measuring semantic similarity among GO terms and gene products. *Bioinformatics*, **26**, 976–978.
14. Uhlén,M., Björling,E., Agaton,C., Szigyarto,C.A., Amini,B., Andersen,E., Andersson,A.C., Angelidou,P., Asplund,A., Asplund,C. *et al.* (2005) A human protein atlas for normal and cancer tissues based on antibody proteomics. *Mol. Cell. Proteomics*, **4**, 1920–1932.
15. Uhlen,M., Oksvold,P., Fagerberg,L., Lundberg,E., Jonasson,K., Forsberg,M., Zwahlen,M., Kampf,C., Wester,K., Hober,S. *et al.* (2010) Towards a knowledge-based Human Protein Atlas. *Nat. Biotechnol.*, **28**, 1248–1250.
16. Gray,K.A., Daugherty,L.C., Gordon,S.M., Seal,R.L., Wright,M.W. and Bruford,E.A. (2013) genenames.org: the HGNC resources in 2013. *Nucleic Acids Res.*, **41**, D545–D552.
17. Munz,M. (2014) *Personalized retrieval of protein-coding disease-relevant gene variants, Institute for Information Systems*. Master's thesis, Technische Universität Braunschweig, Braunschweig.

Downloaded from http://nar.oxfordjournals.org/ at Michigan State University on November 20, 2015